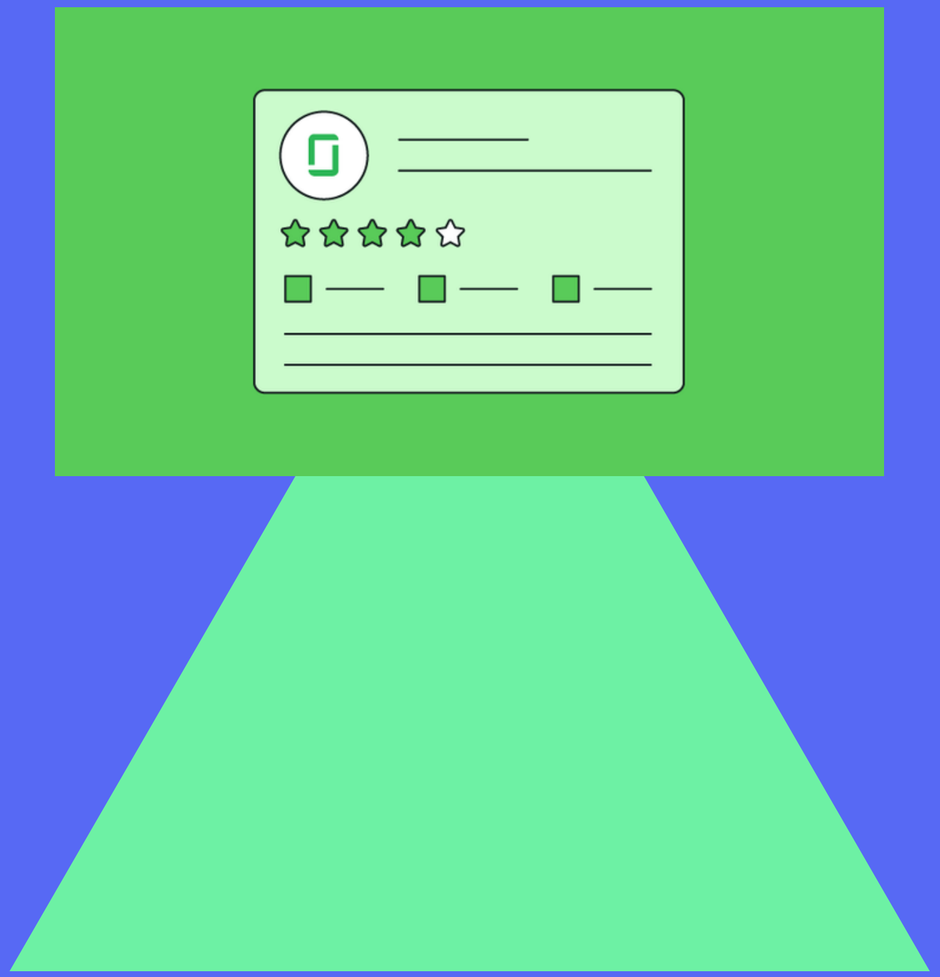
Walking the Walk

Turning Qualitative Text into Quantitative Results



OVERVIEW

The process of finding a job is quite an undertaking with many factors to consider. Ratings from current and former employees can play a key role in whether or not a candidate chooses to apply. In this project, I attempt to close the gap using a translation of qualitative reviews into a quantitative metric that, when combined with machine learning can predict the review score currently available through sites like glassdoor.

Summary of Approach

The 4 Phases

DATA ACQUISITION

Web scraped 14k+ reviews from 150 companies of various industries and ratings from the Glassdoor site.

EDA

During this step non-text and text features was cleaned, recategorized, and visualized..

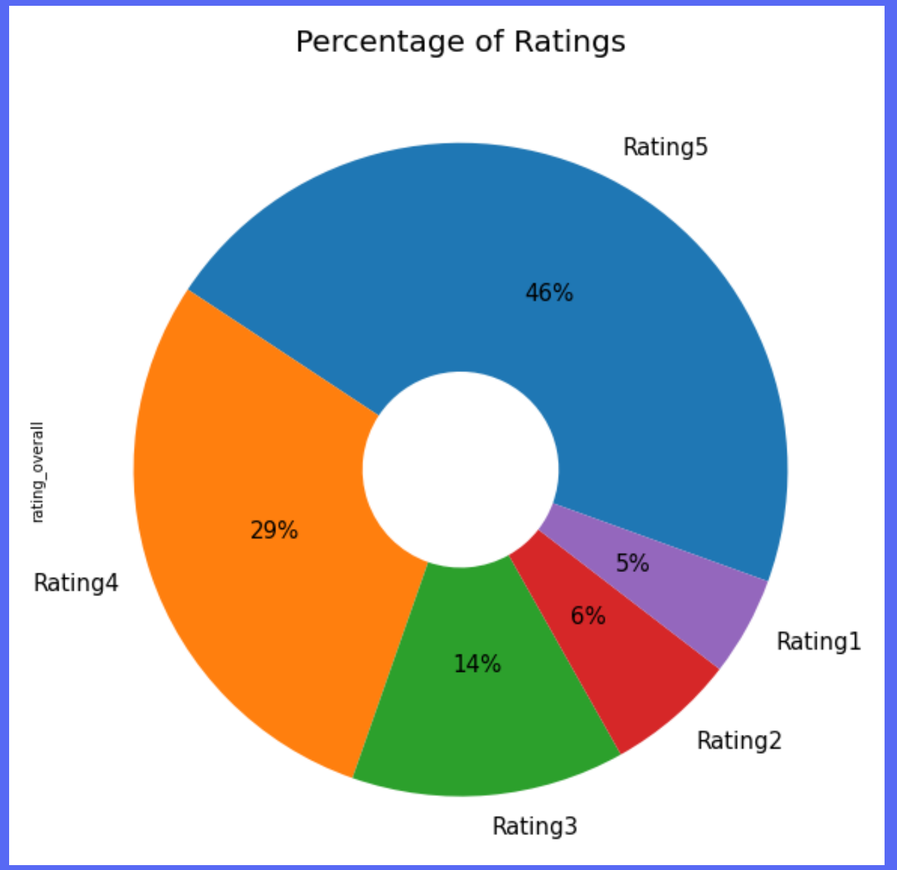
PROCESSING

Prepare the text for modeling through NLP. This involves steps like removing stop words, removing punctuation, stemming, and vectorizing.

MODEL

Laslty, I tested several models in order to find the best fit for my data.

Looking At the Target

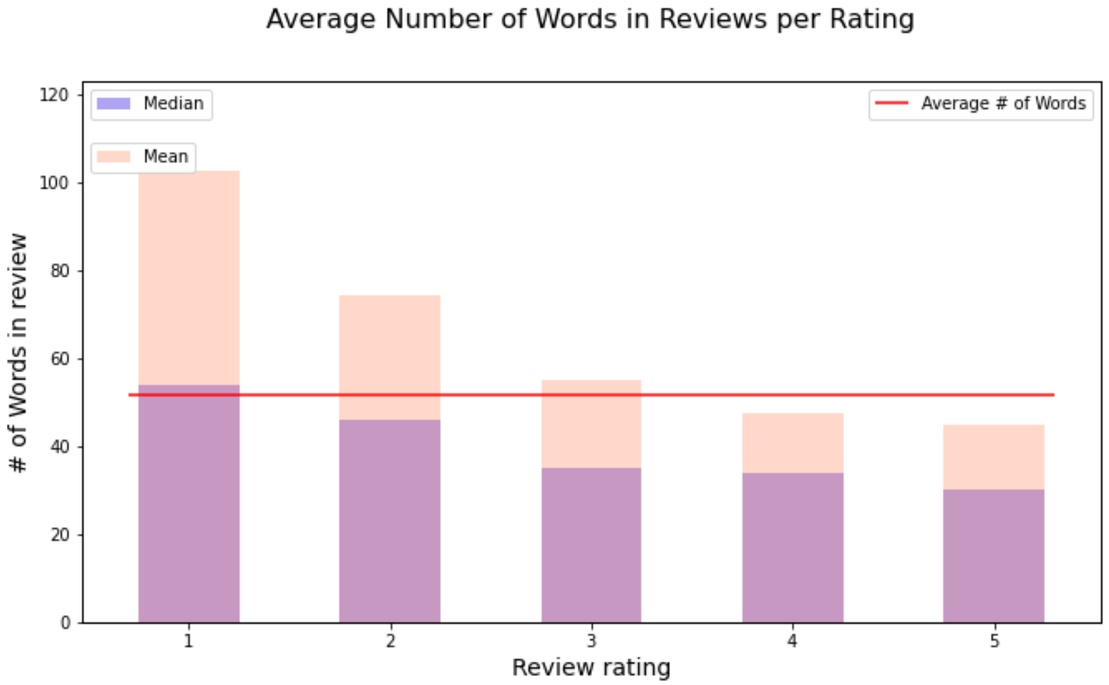


Top 100 Most Common Words

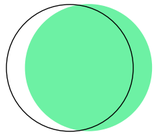
We can see that 'helpfu' is the most popular word throughout all the reviews. We can surmised that reviewers had a lot to say about what they did and did not find helpful at the places of work.

The Worse The Rating...

Another interesting find was that the lower rating the larger the average word count. The worse the experience the more there was to say about it.

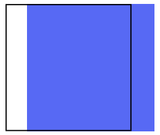


Model Tests



KNN

The KNN Model was the worst performer resulting in an only 45% accuracy.



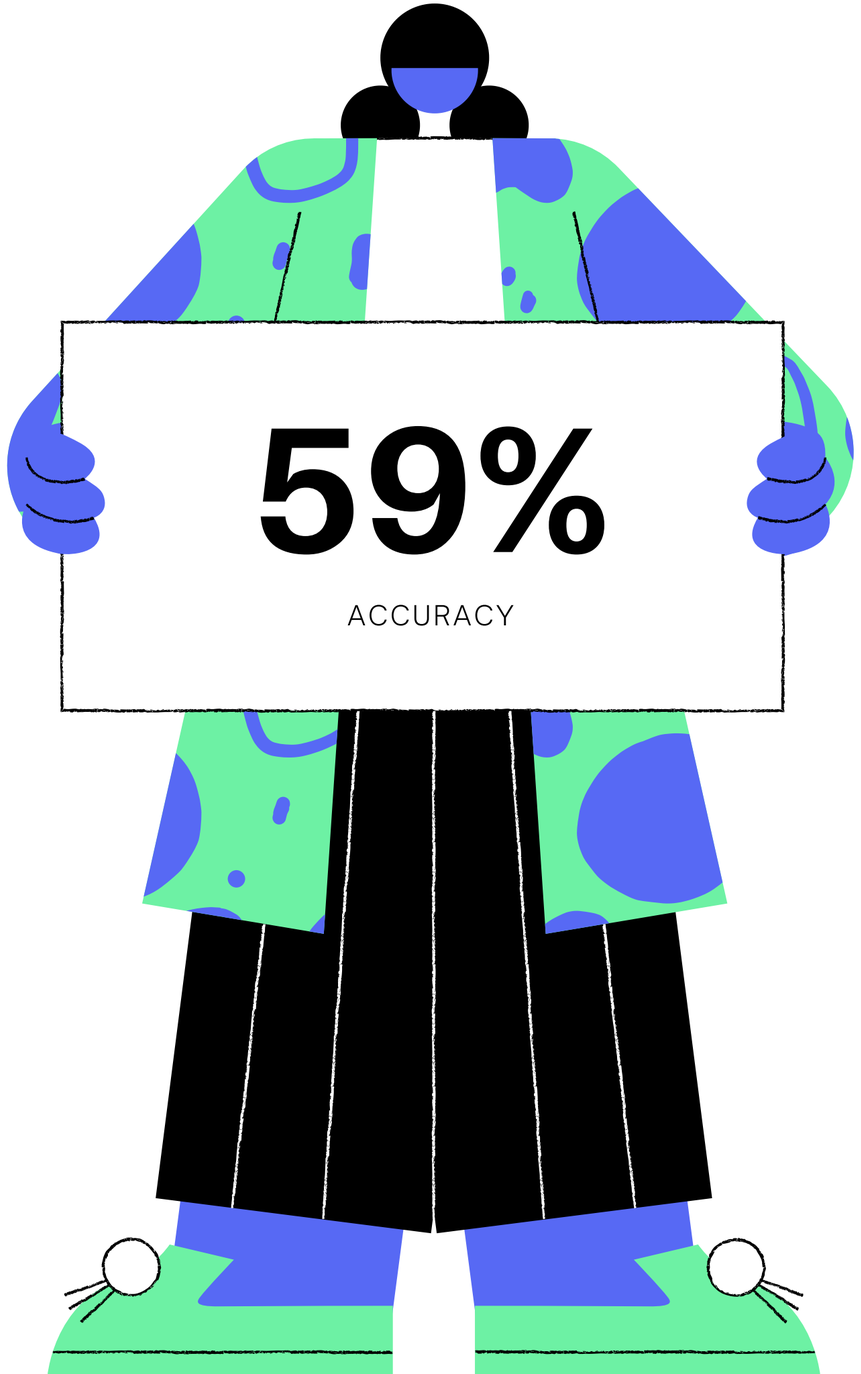
RANDOM FOREST

Random forest performed very near my best performer with 58% accuracy.

LIGHTGBM

Light Gradient Boosted Machine also resulted in a 58% accuracy.

XGBOOST MODEL



Conclusion

This indicates that machine learning could be a better way to score company reviews and this the overall company rating versus relying on a self reported score.

Next Step

For my next steps I will try to improve on my two best models using grid search to see if I can improve accuracy scores.

Contact Me

I'd love to connect.

LINKEDIN

https://www.linkedin.com/in/anjcray/

EMAIL ADDRESS

dataonatangent@gmail.com

GITHUB

https://github.com/DataOnATangent